# Scelta della funzione di legame in un modello lineare generalizzato attraverso delle simulazioni

Fabrizio Bettella

anno accademico 2006/2007

# Indice

# Capitolo 1

# Introduzione

In Statistica, è frequentemente di interesse modellare una variabile risposta in funzione di alcune variabili esplicative. Inoltre, è utile disporre di opportuni test statistici per valutare se le variabili esplicative sono significative, ossia se la loro inclusione nel modello è rilevante. Nel linguaggio statistico questa analisi dei dati viene chiamata analisi di regressione. Il nostro interesse in questa tesi sarà rivolto ai modelli lineari generalizzati. I modelli lineari generalizzati rappresentano un'estensione dei modelli lineari. Alcuni aspetti salienti saranno riassunti nel Capitolo 2. Spesso, diversi modelli lineari generalizzati possono essere considerati ragionevoli per descrivere un assegnato insieme di dati. Ad esempio, si possono considerare diversi insiemi di variabili esplicative, oppure, per un fissato insieme di esplicative, diverse funzioni di legame. Sono allora utili i metodi di selezione del modello. Il più noto tra questi è il criterio di informazione di Akaike (AIC) introdoto da Akaike (1973) e brevemente richiamato nel Capitolo 3. L' obiettivo di questa relazione è valutare tramite uno studio di simulazione le probabilità di selezione corretta dell'AIC qualora si desideri selezionare tra due modelli lineari generalizzati per risposte Poisson, caratterizzati da due

diverse funzioni di legame, con un fissato insieme di variabili esplicative. In particolare, si è interessati a valutare il comportamento dell' AIC all'aumentare del numero di variabili esplicative per una fissata numerosità campionaria. I risultati della simulazione sono presentati nel Capitolo 4, mentre nel Capitolo 5 viene presentato un esempio con dati reali.

# Capitolo 2

# I modelli lineari generalizzati

Un'esempio in cui i modelli lineari generalizzati vengono applicati è quello delle visite mediche in funzione dell'età. E'naturale attendersi che il numero delle visite aumenti con l'età in modo non lineare, quindi un modello lineare risulterebbe inadatto. I modelli lineari si basano sull'ipotesi di normalità dei dati, ma questo spesso non è plausibile per vari motivi. Ad esempio può succedere che la variabilità della variabile risposta cresca in maniera monotona con il suo valore medio oppure potremmo essere interessati a stimare la probabilità che una certa dose di veleno abbia effetto sulle delle cavie. In tutte queste situazioni non è ipotizzabile la distribuzione normale, e questo lo si può capire considerando la distribuzione della variabile risposta.

## 2.1 Un interessante classe di distribuzione

Le principali variabili casuali che fanno parte dei modelli lineari generalizzati sono: la binomiale, la poisson, la gamma e la normale. Esse costituiscono un sottoinsieme della famiglia di

dispersione esponenziale perché la loro funzione di densità può essere scritta come

$$f(y;\theta;\psi) = \exp\left(\frac{\omega}{\psi}\{y\theta - b(\theta)\} + c(y;\psi)\right) \qquad (2.1)$$

dove $\theta, \psi$ sono dei parametri scalari ignoti, $\omega$ è una costante nota, e $b(.)$ e $c(.)$ sono funzioni note la cui scelta individua una particolare distribuzione di probabilità. Quando questo risulta vero possiamo scrivere per una varibile continua $Y$ che

$$Y \sim DE\left(b(\theta), \frac{\psi}{\omega}\right).$$

Per ogni particolare scelta di $\psi$, detto *parametro di dispersione*, la $f(y;\theta)$ costituisce una famiglia esponenziale di parametro $\theta$. Se $\psi$ e $\theta$ variano simultaneamente allora la $f(y;\theta)$ non è una famiglia esponenziale. A parte questa precisazione, nel nostro caso consideriamo $\psi$ come fissato. Per una famiglia esponenziale media e varianza sono date rispettivamente

$$E(y) = b^{'}(\theta)$$

$$Var(y) = b^{''}(\theta)\frac{\psi}{\omega}$$

.

## 2.2 Una nuova classe di modelli

Nei modelli lineari si suppone che le osservazioni $y_i$ siano la realizzazione di una variabile casuale $Y \sim N(\mu_i;\sigma^2)$ e che il *predittore lineare* $\eta_i$ coincida con $x_i^T\beta$ dove $x_i^T$ è la i-esima riga di $X$ per $i = 1,\ldots,n$. Inoltre la relazione tra valore medio $\mu_i$ e il predittore lineare $\eta_i$ è l'identità. Schematizzando avremo $Y \sim N(\mu_i;\sigma^2)$ $\mu_i = \eta_i \quad \eta_i = x_i^T\beta$. I modelli lineari generalizzati rappresentano un'estensione dei modelli lineari per due motivi

1. come distribuzione possibile per $Y_i$ non considerano solo la normale, ma qualunque altra distribuzione $DE\left(b(\theta_i); \frac{\psi}{\omega_i}\right)$ tale che $b^{'}(\theta_i) = \mu_i$
2. ipotizzano varie forme di legame fra il valore medio e il predittore lineare, cioè avremo che $g(\mu_i) = \eta_i$ dove $g(.)$ è una funzione monotona derivabile detta *funzione di legame*.

In altre parole diremo che i modelli lineari generalizzati sono caratterizzati dai seguenti elementi $Y_i \sim DE\left(b(\theta_i); \frac{\psi}{\omega_i}\right)$ con $b^{'}(\theta_i) = \mu_i$ struttura stocastica, $g(\mu_i) = \eta_i$ funzione legame, $\eta_i = x_i^T\beta$ predittore lineare.

## 2.3 Stima dei parametri

Date le osservazioni campionarie $y_1, \ldots, y_n$ vogliamo fare inferenza sul parametro $\beta$ . Il nostro interesse è rivolto a $\beta$ perchè determina la relazione tra le variabili esplicative e $\mu$, mentre $\psi$ è un parametro di disturbo quando è presente. Indichiamo con $p$ la dimensione di $\beta$ e con $X = (x_{ij})$ la matrice $n \times p$ con i-esima riga $x_i^T$. Essendo le osservazioni indipendenti, la log-verosimiglianza sarà

$$\ell(\beta) = \sum_{i=1}^{n} \left( \frac{\omega_i(y_i\theta_i - b(\theta_i))}{\psi} + c_i(y_i; \psi) \right) = \sum_{i=1}^{n} \ell_i(\beta).$$

Per facilitare il calcolo delle equazioni di verosimiglianza usiamo

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\eta_i}{\partial \beta_j} \tag{2.2}$$

i cui termini possono essere riscritti nel seguente modo

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{\psi/\omega_i} = \frac{y_i - \mu_i}{\psi/\omega_i}, \tag{2.3}$$

$$\frac{\partial \mu_i}{\partial \theta_i} = b^{''}(\theta_i) = \frac{\omega_i var(Y_i)}{\psi}, \tag{2.4}$$

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}. \tag{2.5}$$

Quindi abbiamo

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{y_i - \mu_i}{\psi/\omega_i} \frac{\psi/\omega_i}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij}$$

A questo punto le equazioni di verosimiglianza per $\beta$ hanno la seguente forma

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \quad (j = 1, \ldots, p). \tag{2.6}$$

# Capitolo 3

# Criteri di selezione del modello

Nella letteratura statistica esistono vari metodi o criteri per scegliere un modello tra quelli plausibili. Il piu' noto è senza dubbio l'AIC, *Akaike Information Criterion*, ma ve ne sono altri come ad esempio il BIC, *Bayesian Information Criterion*. Il criterio di Akaike viene calcolato nel seguente modo

$$-2\log\left(L\left(\hat{\theta};y\right)\right)+2p$$

mentre quello Bayesiano è dato da

$$p\log\left(n\right)-2\log\left(L\left(\hat{\theta};y\right)\right).$$

In entrambi i casi il modello scelto sarà quello che minimizza il criterio. Con il metodo di Akaike si tende a selezionare un modello leggermente sovraparametrizzato per $n$ sufficientemente elevato, mentre con quello Bayesiano per $n$ non elevato si tende a selezionarne uno leggermente sottoparametrizzato. Come detto in precedenza, faremo riferimento all'AIC. L'obbiettivo è di scegliere un modello intermedio tra tutti quelli correttamente specificati, che bilanci bontà di adattamento ai dati e parsimonia della modellazione. Un modello con molti parametri è piu'

flessibile e permette una migliore rappresentazione dei dati. Tutto questo è vero fino ad un certo punto, perchè se il modello è troppo ampio può succedere che le stime dei parametri tendano a divenire instabili, cioè gli errori quadratici medi stimati tendono a gonfiarsi. Per generalizzare, i modelli parametrici che si prendono in considerazione sono annidati, cioè avremo

$$\Theta_1 \subset \Theta_2 \subset \ldots \subset \Theta_k \subseteq \Re^k$$

dove $\Theta_i$ per $i = 1, \ldots, k$ rappresenta lo spazio parametrico per un generico modello. Indicando con $\theta^{(1)}, \ldots, \theta^{(k)}$ i parametri dei vari modelli, il passaggio da $\Theta_k = \left\{\theta^{(k)} = (\theta_1, \ldots, \theta_k)\right\}$ a $\Theta_{k-1}$ avviene attraverso l'ipotesi $\theta_k = 0$. Aggiungendo ulteriori annullamenti alle componenti otteniamo i corrispondenti modelli di complessità inferiore. I criteri per la selezione di un modello si basano sulle verosimiglianze profilo penalizzate per il numero di parametri presenti. Questa penalizzazione per la dimensione parametrica, nei principali metodi viene basata su considerazioni di informazione predittiva del modello, cioè sulla perdita attesa di efficacia predittiva. I criteri di Akaike e quello Bayesiano derivano proprio da queste considerazioni.

# Capitolo 4

# Studio di simulazione

In questo capitolo mostreremo i risulati di alcune simulazioni fatte con R, cioè replicheremo varie numerosità da una disribuzione di Poisson prima con vera media $\exp(\beta_1 x_{ik} + \ldots + \beta_k x_{ik})$ e poi con vera media $(\beta_1 x_{ik} + \ldots + \beta_k x_{ik})^2$. In entrambe le situazioni, verrà confrontato l'AIC ottenuto con legame canonico e quello ottenuto con legame radice quadrata. Le numerosità campionarie scelte per questo nostro esperimento sono $n = 5, 10, 20, 50$ e le replicazioni per ogni campione sono 5000. Queste simulazioni prevedono vari casi:

- valori delle esplicative compresi fra 0 e 1 e valori maggiori di 1;
- due regressori, compresa l'intercetta, entrambi positivi,
- due regressori, compresa l'intercetta, con $\beta_2$ minore di 0;
- tre regressori includendo l'intercetta;
- quattro regressori includendo l'intercetta;

Nei casi con due parametri, $\beta_1$ e $\beta_2$, faremo un grafico delle medie per dare un' idea del loro andamento. Ovviamente, i casi analizzati avranno le due medie con lo stesso andamento.

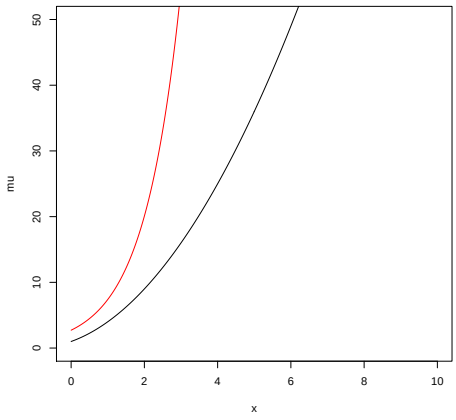


Figura 4.1: Grafico medie

## 4.1 Caso con due regressori positivi

Nel seguente esempio le due medie in questione sono $\exp(1 + x_{i2})$ e $(1 + x_{i2})^2$. Un grafico per le medie è dato dalla Figura 4.1. Osservandolo si nota che la Poisson con media esponenziale cresce piu' rapidamente rispetto all'altra. Per valori di $x$ maggiori di 1 avremo le tabelle dalla 4.1 fino alla 4.4, mentre per valori di $x$ minori di 1 le tabelle dalla 4.5 fino alla 4.8.

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.536 | 0.3058 |
| Link=sqrt | 0.464 | 0.6942 |

Tabella 4.1: Tabella con n=5

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.524 | 0.0384 |
| Link=sqrt | 0.476 | 0.9616 |

Tabella 4.2: Tabella con n=10

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.8018 | 0.0152 |
| Link=sqrt | 0.1982 | 0.9848 |

Tabella 4.3: Tabella con n=20

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 1 | 0 |
| Link=sqrt | 0 | 1 |

Tabella 4.4: Tabella con n=50

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.4648 | 0.4822 |
| Link=sqrt | 0.5352 | 0.5178 |

Tabella 4.5: Tabella con n=5

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.476 | 0.4758 |
| Link=sqrt | 0.524 | 0.5242 |

Tabella 4.6: Tabella con n=10

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.4906 | 0.454 |
| Link=sqrt | 0.5094 | 0.546 |

Tabella 4.7: Tabella con n=20

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.5058 | 0.464 |
| Link=sqrt | 0.4942 | 0.536 |

Tabella 4.8: Tabella con n=50

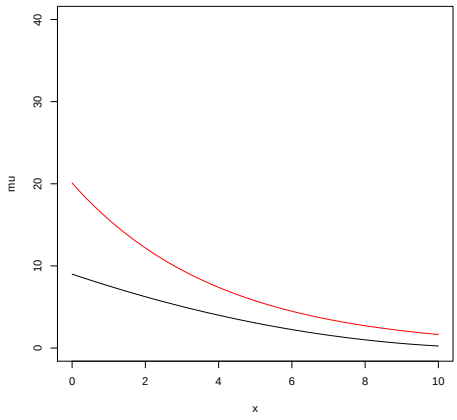


Figura 4.2: Grafico medie

## 4.2 Caso con $\beta_2$ minore di 0

Per questo caso, abbiamo scelto le medie $\exp(3 - 0.25x_{i2})$ e $(3 - 0.25x_{i2})^2$. Una rappresentazione grafica è data dalla Figura 4.2. Dalla figura vediamo che la Poisson con media esponenziale decresce un po' piu' rapidamente rispetto all'altra. Le tabelle dalla 4.9 fino alla 4.12 si riferiscono a valori delle esplicative maggiori di 1 mentre dalla 4.13 alla 4.16 per valori minori di 1.

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.5062 | 0.4606 |
| Link=sqrt | 0.4938 | 0.5394 |

Tabella 4.9: Tabella con n=5

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.5132 | 0.467 |
| Link=sqrt | 0.4868 | 0.533 |

Tabella 4.10: Tabella con n=10

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.5182 | 0.4588 |
| Link=sqrt | 0.4818 | 0.5412 |

Tabella 4.11: Tabella con n=20

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.5468 | 0.4654 |
| Link=sqrt | 0.4532 | 0.5346 |

Tabella 4.12: Tabella con n=50

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.4942 | 0.5016 |
| Link=sqrt | 0.5058 | 0.4984 |

Tabella 4.13: Tabella con n=5

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.4966 | 0.4846 |
| Link=sqrt | 0.5034 | 0.5154 |

Tabella 4.14: Tabella con n=10

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.5044 | 0.4822 |
| Link=sqrt | 0.4956 | 0.5178 |

Tabella 4.15: Tabella con n=20

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.4978 | 0.4868 |
| Link=sqrt | 0.5022 | 0.5132 |

Tabella 4.16: Tabella con n=50

## 4.3 Caso con tre regressori

Le funzioni delle medie in questione sono $\exp(1 + 0.01x_{i2} + 0.05x_{i3})$ e $(1 + 0.01x_{i2} + 0.05x_{i3})^2$. In questo esempio non faremo un grafico perchè non sarebbe di grande utilità. Per valori delle esplicative maggiori di 1 abbiamo le tabelle dalla 4.17 alla 4.20, mentre per valori minori di1 avremo quelle dalla 4.21 alla 4.24.

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.4742 | 0.4958 |
| Link=sqrt | 0.5258 | 0.5042 |

Tabella 4.17: Tabella con n=5

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.4684 | 0.4796 |
| Link=sqrt | 0.5316 | 0.5204 |

Tabella 4.18: Tabella con n=10

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.4612 | 0.484 |
| Link=sqrt | 0.5388 | 0.516 |

Tabella 4.19: Tabella con n=20

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.4832 | 0.4836 |
| Link=sqrt | 0.5168 | 0.5164 |

Tabella 4.20: Tabella con n=50

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.4672 | — |
| Link=sqrt | 0.528 | — |

Tabella 4.21: Tabella con n=5

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.5202 | — |
| Link=sqrt | 0.4798 | — |

Tabella 4.22: Tabella con n=10

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.478 | 0.4664 |
| Link=sqrt | 0.522 | 0.5336 |

Tabella 4.23: Tabella con n=20

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.486 | 0.4742 |
| Link=sqrt | 0.514 | 0.5258 |

Tabella 4.24: Tabella con n=50

## 4.4 Caso con quattro regressori

Le medie in questione sono date da $\exp(1 + 0.2x_{i2} - 0.15x_{i3} + 0.3x_{i4})$ e $(1 + 0.2x_{i2} - 0.15x_{i3} + 0.3_{i4})^2$. Con valori delle esplicative maggiori di 1 avremo tabelle dalla 4.25 alla 4.28, mentre con valori minori di 1 le tabelle dalla 4.29 alla 4.32. [h]

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.4824 | 0.5 |
| Link=sqrt | 0.5176 | 0.5 |

Tabella 4.25: Tabella con n=5

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.4744 | 0.4616 |
| Link=sqrt | 0.5256 | 0.5384 |

Tabella 4.26: Tabella con n=10

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.4888 | 0.4708 |
| Link=sqrt | 0.5112 | 0.5292 |

Tabella 4.27: Tabella con n=20

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.4942 | 0.4694 |
| Link=sqrt | 0.5058 | 0.5306 |

Tabella 4.28: Tabella con n=50

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.4642 | — |
| Link=sqrt | 0.5358 | — |

Tabella 4.29: Tabella con n=5

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.4628 | 0.457 |
| Link=sqrt | 0.5372 | 0.543 |

Tabella 4.30: Tabella con n=10

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.4742 | 0.4924 |
| Link=sqrt | 0.5258 | 0.5076 |

Tabella 4.31: Tabella con n=20

| Modello gen./Modello scelto | Link=log | Link=sqrt |
|---|---|---|
| Link=log | 0.4838 | 0.4778 |
| Link=sqrt | 0.5162 | 0.5222 |

Tabella 4.32: Tabella con n=50

Dalle simulazioni effettuate emerge quanto segue:

- nel caso con due regressori, entrambi positivi e per valori di $x$ maggiori di 1, all' aumentare della numerosità campionaria aumenta anche in maniera significativa la probabilità di selezione del modello generatore;
- nel caso di due regressori con $\beta_2$ minore di 0 la probabilità di selezione del modello generatore varia attorno al 50 per cento, sia per valori di $x$ maggiori di 1 che per valori compresi fra 0 e 1;
- con tre regressori, entrambi positivi, la probabilità di selezione del modello generatore diminuisce un po' rispetto al caso precedente, anche se siamo sempre sul 50 per cento, sia per valori delle $x$ maggiori di 1 che per valori compresi fra 0 e 1;
- nel caso di quattro regressori, con il solo $\beta_3$ minore di 0, si registra una ulteriore diminuzione, anche se non elevata, della probabilità di selezione del modello generatore sia per valori delle $x$ maggiori di 1 che per valori compresi fra 0 e 1.

In definitiva dai risultati ottenuti vediamo che il criterio dell'AIC, all'aumentare del numero delle varibili esplicative, potrebbe non essere attendibile.

# Capitolo 5

# Un esempio con dati reali

L'insieme dei dati che useremo contiene variabili che sono state rilevate per studiare se la quantità di silicio presente nei polmoni dipende dalle variabili esplicative rilevate. Con questi dati ci proponiamo di confermare quanto visto precedentemente con le simulazioni e quindi non faremo affermazioni riguardo le variabili in questione. L'insieme dei dati è il seguente:

| | group | Age | Gender | Smoke | n.positive | n.spot |
|---|---|---|---|---|---|---|
| 1 | exposed | 60 | M | yes | 31 | 236 |
| 2 | exposed | 49 | M | yes | 24 | 184 |
| 3 | control | 66 | F | no | 3 | 5 |
| 4 | control | 43 | M | yes | 4 | 7 |
| 5 | control | 67 | M | yes | 4 | 6 |
| 6 | control | 72 | M | no | 7 | 13 |
| 7 | control | 57 | M | no | 13 | 14 |
| 8 | control | 65 | M | no | 5 | 5 |
| 9 | control | 66 | F | no | 4 | 5 |
| 10 | control | 70 | M | yes | 7 | 12 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 11 | control | 69 | F | yes | 5 | 5 |
| 12 | control | 48 | M | no | 6 | 18 |
| 13 | control | 76 | F | yes | 8 | 12 |
| 14 | control | 72 | F | no | 5 | 13 |
| 15 | control | 70 | F | yes | 0 | 0 |
| 16 | control | 54 | M | yes | 0 | 0 |
| 17 | control | 47 | F | yes | 0 | 0 |
| 18 | control | 50 | M | no | 0 | 0 |
| 19 | control | 52 | F | no | 0 | 0 |
| 20 | control | 66 | M | no | 0 | 0 |
| 21 | control | 71 | F | yes | 0 | 0 |
| 22 | control | 68 | M | no | 0 | 0 |
| 23 | control | 69 | M | no | 0 | 0 |
| 24 | control | 49 | F | yes | 0 | 0 |
| 25 | control | 77 | F | no | 0 | 0 |
| 26 | control | 66 | F | no | 0 | 0 |
| 27 | control | 62 | M | yes | 0 | 0 |
| 28 | normal | 76 | F | no | 6 | 6 |
| 29 | normal | 65 | M | yes | 4 | 8 |
| 30 | normal | 74 | M | no | 6 | 11 |
| 31 | normal | 50 | F | no | 0 | 0 |
| 32 | normal | 82 | F | no | 0 | 0 |
| 33 | normal | 70 | M | no | 0 | 0 |
| 34 | normal | 16 | F | no | 0 | 0 |
| 35 | normal | 61 | M | yes | 0 | 0 |
| 36 | normal | 82 | F | no | 0 | 0 |
| 37 | normal | 66 | M | no | 0 | 0 |
| 38 | normal | 61 | M | no | 0 | 0 |
| 39 | normal | 48 | F | no | 0 | 0 |
| 40 | normal | 68 | M | no | 0 | 0 |
| 41 | normal | 75 | M | no | 0 | 0 |

```
42  normal  68      M   yes          0      0
43  normal  80      M    no          0      0
44  normal  47      M   yes          0      0
45  normal  67      M    no          0      0
46  normal  38      F   yes          0      0
47  normal  55      F    no          0      0
```

Il numero di pazienti è $n = 47$ e le variabili esplicative rilevate su di essi sono:

- 'group':fattore a tre livelli dove exposed rappresenta i decessi per tumore al polmone ed esposizione nel lavoro al silicio,control i decessi per tumore al polmone per cause ignote e normal i decessi per cause ignote;
- 'age': l'età dei pazienti;
- 'gender': fattore che indica il genere dei soggetti;
- 'smoke': fattore che ci dice se il soggetto era fumatore oppure no;
- 'n.positive': rappresenta il numero di zone del polmone contenenti tracce di silicio;
- 'n.spot': indica la quantità di silicio trovata nelle zone del polmone ritenute positive.

Si desidera valutare la variabile risposta 'n.spot' (di conteggio), quindi un modello adeguato puo' essere una regressione poissoniana. Si adotteranno sia la funzione legame canonica sia quella radice quadrata. La sintassi da usare in R nel primo caso è

```
fit1<-glm(n.spot~group+Age+Gender+Smoke+n.positive,poisson(log))
```

mentre nel secondo avremo

```
fit2<-glm(n.spot~group+Age+Gender+Smoke+n.positive,poisson(sqrt)).
```

I valori nei rispettivi casi dell'AIC sono 295.3 e 135.9. Sapendo che un piu' basso valore dell'AIC corrisponde a una maggiore verosimiglianza, in questo caso risulterebbe migliore il legame *sqrt*. Guardando le simulazioni fatte in precedenza però, notiamo che all'aumentare dei regressori la probabilità di selezione fra legame canonico e legame radice quadrata tende a diminuire, e questo si verifica anche considerando valori delle esplicative sia maggiori che minori di 1. Un incremento della numerosità campionaria nei vari esempi non porta a cambiamenti significativi, eccetto nel primo caso per valori di $x$ maggiori di 1. Da ciò si deduce che il criterio appena citato poterbbe essere non adatto per valutare quale legame da utilizzare sia migliore, anche se 135.9 è nettamente inferiore a 295.3.

# Bibliografia

[1] Azzalini,A.(2004). *Inferenza Statistica. Una presentazione basata sul concetto di verosimiglianza* Springer-Verlag, Milano.

[2] Pace, L. e Salvan, A. (2001). *Introduzione alla Statistica II - Inferenza,verosimiglianza, modelli.* Cedam.